PCT/EP2004/051266
14. 07. 2004

BUNDESREPUBLIK DEUTSCHLAND

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

Prioritätsbescheinigung über die Einreichung einer Patentanmeldung

Aktenzeichen: 103 30 280.8

Anmeldetag: 4. Juli 2003

Anmelder/Inhaber: Siemens Aktiengesellschaft, 80333 München/DE

Bezeichnung: Verfahren, Computerprogramm mit Programmcode-Mitteln und Computerprogramm-Produkt zur Anlayse eines regulatorischen genetischen Netzwerks einer Zelle

IPC: C 12 Q 1/68

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 9. Juli 2004
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Stanschus

A 9161
06/00
EDV-L

Beschreibung

Verfahren, Computerprogramm mit Programmcode-Mitteln und Computerprogramm-Produkt zur Analyse eines regulatorischen genetischen Netzwerks einer Zelle

Die Erfindung betrifft eine Analyse eines regulatorischen genetischen Netzwerks einer Zelle unter Verwendung eines statistischen Verfahrens.

Aus [1] sind Grundlagen eines regulatorischen genetischen Netzwerks einer Zelle bekannt. Unter einem solchen regulatorischen genetischen Netzwerk seien dabei im Folgenden insbesondere regulatorische Wechselwirkungen zwischen Genen einer Zelle verstanden.

Ein Genom, d.h. die menschliche Erbsubstanz, umfasst schätzungsweise 20.000 bis 40.000 Gene, von denen jeweils eine biologisch bestimmte Anzahl - abhängig von einer Spezialisierung einer Zelle - in Form einer DNA oder eines Teils einer DNA in einer Zelle vorhanden sind.

Als ein Gen wird dabei ein nicht notwendigerweise zusammenhängender Abschnitt dieser DNA bezeichnet, der einen genetischen Code für ein Protein oder auch für eine Gruppe von Proteinen (Eiweißstoffe) bzw. für eine Erzeugung eines Proteins oder einer Proteingruppe enthält. Insgesamt beinhalten die Gene einen genetischen Code für etwa eine Million Proteine.

Ein Wechselspiel bzw. die Wechselwirkungen der Gene untereinander sowie mit den Proteinen stellt den wichtigsten Teil einer Maschinerie (regulatorisches genetisches Netzwerk) dar, die einer Entwicklung eines menschlichen Körpers aus einer befruchteten Eizelle sowie allen Körperfunktionen zugrunde liegt.

Auch aus [1] ist bekannt, dass sogenannte Gen-Expressionsraten, welche ein Gen-Expressionsmuster bilden, eine Beschreibung bzw. Repräsentation eines regulatorischen genetischen Netzwerks bzw. eines aktuellen Zustands des regulatorischen genetischen Netzwerks liefern.

Vereinfacht oder anschaulich ausgedrückt repräsentiert somit ein Gen-Expressionsmuster einer Zelle einen Zustand des regulatorischen genetischen Netzwerks dieser Zelle.

Ferner ist bekannt, dass unter Verwendung von Hochdurchsatz-Genexpressions-Messungen (Microarray-Daten) diese Gen-Expressionsraten messbar sind. Die Microarray-Daten beschreiben wiederum Momentaufnahmen des Gen-Expressionsmusters.

Viele Krankheiten und Fehlfunktionen des Körpers gehen auf Störungen des regulatorischen genetischen Netzwerks zurück, welche sich in eine stark veränderten Gen-Expressionsverhalten (Gen-Expressionsraten) bzw. einem veränderten Gen-Expressmuster einer Zelle widerspiegeln.

Somit stellt ein Verständnis des regulierenden genetischen Netzwerks einen wichtigen Schritt auf dem Weg zu einer Charakterisierung und einem Verstehen von genetischen Mechanismen sowie in weiterer Folge zu einer Identifizierung von sogenannten dominanten oder Funktionsstörungen auslösenden Genen dar, welche den Krankheiten oder Fehlfunktionen zugrunde liegen.

Beispielsweise kann in einer Krebsforschung, bei der die Identifizierung von Geschwülste und Tumore unterdrückenden Genen eine Schlüsselrolle spielt, die Kenntnis neuer potenzieller Onkogene und ihre Wechselwirkung mit anderen Genen ein Beitrag zu einer Aufdeckung von Grundprinzipien (von Krebserkrankungen) sein, welche ein Umwandlung normaler Zellen in bösartige Krebszellen bestimmen.

Weitergehend ist für eine Entwicklung von verbesserten Medikamenten und Therapien zur Bekämpfung von genetischen Krankheiten daher ebenfalls ein quantitatives Verständnis des regulatorischen genetischen Netzwerks einer Zelle erforderlich.

So wirken einige Medikamente als Agonisten bzw. Antagonisten spezifischer Zielproteine, d. h. sie verstärken oder schwächen die Funktion eines Proteins mit entsprechender Rückwirkung auf das regulatorische genetische Netzwerk mit dem Ziel, dieses zurück in einen normalen Funktionsmodus zu bringen.

Aus [2] ist eine Beschreibung eines regulatorischen genetischen Netzwerks einer Zelle unter Verwendung eines statistischen Verfahrens, eines kausalen Netzes, bekannt.

Aus [3] ist ein kausales Netz, ein Bayesianisches (Bayessches) Netzwerk, bekannt.

Bayessche Netzwerke

Ein Bayessches Netzwerk *B* ist ein spezieller Typ der Darstellung einer gemeinsamen multivariaten Wahrscheinlichkeitsdichtefunktion (WDF) einer Menge von Variablen X durch ein graphisches Modell.

Es ist durch einen gerichteten azyklischen Graphen (directed acyclic graph, *DAG*) *G* definiert, in welchem jeder Knoten i = 1, ..., n einer Zufallsvariablen X_i entspricht.

Die Kanten zwischen den Knoten repräsentieren statistische Abhängigkeiten und können als Kausalzusammenhänge zwischen ihnen interpretiert werden. Der zweite Bestandteil des Bay-

esschen Netzwerkes ist die Menge von bedingten WDFen $P(X_i|Pa_i, \theta, G)$, welche mittels eines Vektors θ parametriert sind.

Diese bedingten WDFen spezifizieren die Art der Abhängigkeiten der einzelnen Variablen *i* von der Menge ihrer Elternknoten (Parents) Pa_i. Somit kann die gemeinsame WDF in die Produktform

(1) $$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i|Pa_i, \Theta, G)$$

zerlegt werden.

Der *DAG* eines Bayesschen Netzwerkes beschreibt auf eindeutige Weise die bedingten Abhängigkeits- und Unabhängigkeitsbeziehungen zwischen einer Menge von Variablen, jedoch hat im Gegensatz dazu eine gegebene statistische Struktur der WDF keinen eindeutigen *DAG* zur Folge.

Vielmehr kann gezeigt werden, dass zwei *DAG* ein und dieselbe WDF beschreiben, dann und nur dann, wenn sie dieselbe Menge von Kanten und dieselbe Menge von "Colliders" aufweisen, wobei ein Collider eine Konstellation ist, in welcher wenigstens zwei gerichtete Kanten zu demselben Knoten führen.

Der Erfindung liegt die Aufgabe zugrunde, ein Verfahren anzugeben, welches eine Analyse eines regulatorischen genetischen Netzwerks einer Zelle, beispielsweise repräsentiert durch ein Gen-Expressionsmuster der Zelle, ermöglicht.

Ferner liegt der Erfindung die Aufgabe zugrunde, ein Verfahren anzugeben, welches eine Identifikation eines defekten

Gens, beispielsweise eines Onko- oder Tumor-Gens, in dem regulatorischen genetischen Netzwerk einer Zelle ermöglicht.

Weiter soll die Erfindung eine Simulation und/oder eine Analyse einer Wirkweise eines Medikaments auf das regulatorische genetische Netzwerk einer Zelle ermöglichen.

Diese Aufgabe wird durch das Verfahren, durch das Computerprogramm mit Programmcode-Mitteln und das Computerprogramm-Produkt zur Analyse eines regulatorischen genetischen Netzwerks einer Zelle mit den Merkmalen gemäß dem jeweiligen unabhängigen Patentanspruch gelöst.

Bei dem grundlegenden Verfahren zur Analyse eines regulatorischen genetischen Netzwerks einer Zelle wird ein kausales Netz verwendet,

- welches kausale Netz das regulatorische genetische Netzwerk der Zelle beschreibt derart, dass Knoten des kausalen Netzes Gene des regulatorischen genetischen Netzwerks repräsentieren und Kanten des kausalen Netzes regulatorische Wechselwirkungen zwischen den Genen des regulatorischen genetischen Netzwerks repräsentieren.

Bei dem Analyseverfahren wird nun für ein ausgewähltes Gen des regulatorischen genetischen Netzwerks eine Gen-Expressionsrate vorgegeben. Unter Verwendung des kausalen Netzes wird für die vorgegebene Gen-Expressionsrate ein resultierendes Gen-Expressionsmuster für das regulatorische genetische Netzwerk generiert. Das generierte resultierende Gen-Expressionsmuster wird anschließend mit einem vorgegebenen Gen-Expressionsmuster des regulatorischen genetischen Netzwerks verglichen.

Das Computerprogramm mit Programmcode-Mitteln ist eingerichtet, um alle Schritte gemäß dem erfindungsgemäßen Verfahren

durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

Das Computerprogramm-Produkt mit auf einem maschinenlesbaren Träger gespeicherten Programmcode-Mitteln ist eingerichtet, um alle Schritte gemäß dem erfindungsgemäßen Verfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

Die Anordnung sowie das Computerprogramm mit Programmcode-Mitteln, eingerichtet um alle Schritte gemäß dem erfinderischen Verfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird, sowie das Computerprogramm-Produkt mit auf einem maschinenlesbaren Träger gespeicherten Programmcode-Mitteln, eingerichtet um alle Schritte gemäß dem erfinderischen Verfahren durchzuführen, wenn das Programm auf einem Computer ausgeführt wird, sind insbesondere geeignet zur Durchführung des erfindungsgemäßen Verfahrens oder einer seiner nachfolgend erläuterten Weiterbildungen.

Eine probabilistische Semantik eines kausalen Netzes, wie eines Bayesschen Netzwerkes, ist zur Analyse von Gen-Expressionsraten, beispielsweise gegeben in Form von Microarray-Daten, sehr gut geeignet, da sie an die stochastische Natur sowohl von biologischen Prozesse als auch von mit einem Rauschen behafteten Experimente angepasst ist.

Ferner wird, anschaulich gesehen, ein Effekt eines Expressionszustandes bestimmter Gene auf ein globales Gen-Expressionsmuster (inverse Modellierung) geschätzt, indem ein resultierendes Gen-Expressionsmuster analysiert wird.

Bevorzugte Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Die im weiteren beschriebenen Weiterbildungen beziehen sich sowohl auf die Verfahren als auch auf die Anordnung.

Die Erfindung und die im weiteren beschriebenen Weiterbildungen können sowohl in Software als auch in Hardware, beispielsweise unter Verwendung einer speziellen elektrischen Schaltung, realisiert werden.

Ferner ist eine Realisierung der Erfindung oder einer im weiteren beschriebenen Weiterbildung möglich durch ein computerlesbares Speichermedium, auf welchem das Computerprogramm mit Programmcode-Mitteln gespeichert ist, welches die Erfindung oder Weiterbildung ausführt.

Auch kann die Erfindung oder jede im weiteren beschriebene Weiterbildung durch ein Computerprogrammerzeugnis realisiert sein, welches ein Speichermedium aufweist, auf welchem das Computerprogramm mit Programmcode-Mitteln gespeichert ist, welches die Erfindung oder Weiterbildung ausführt.

Bei einer Weiterbildung wird das ausgewählte Gen unter Verwendung des kausalen Netzes mittels einer Abhängigkeitsanalyse ausgewählt.

Auch kann die Gen-Expressionsrate des ausgewählten Genes derart vorgegeben werden, dass die vorgegebene Gen-Expressrate des ausgewählten Genes eine Annahme eines Gendefekts widerspiegelt.

Als kausales Netz kann ein Bayesianisches bzw. Bayessches Netz verwendet werden.

Auch kann das kausale Netz von einem Typ DAG (directed acylic graph) sein.

Ferner kann bzw. können das generierte resultierende und/oder das vorgegebene Gen-Expressionsmuster diskrete Genzustände repräsentieren, wobei die repräsentierten diskreten Genzustände ein über-, ein normal-, ein unterexprimierten Genzustand sein können.

Bei einer Weiterbildung wird der Vergleich des generierten resultierenden Gen-Expressionsmuster mit dem vorgegebenen Gen-Expressionsmuster unter Verwendung eines statischen Verfahrens und/oder einer statistischen Kennzahl, insbesondere eines Abstandsmaßes, durchgeführt.

Auch kann vorgesehen werden, dass das kausale Netz unter Verwendung von Gen-Expressionsmustern trainiert wird, wobei die Knoten und die Kanten des kausalen Netzes angepasst werden.

Ferner ist es zweckmäßig, dass die Gen-Expressionsmuster, insbesondere das vorgegebene Gen-Expressionsmuster und/oder die Gen-Expressionsmuster für das Training, bestimmt werden unter Verwendung einer DNA-Micro-Array-Technik.

Bei einer Ausgestaltung ist das vorgegebene Gen-Expressionsmuster und/oder die Gen-Expressionsmuster für das Training ein Gen-Expressionsmuster eines genetischen regulatorischen Netzwerks einer kranken Zelle.

Dabei kann beispielsweise die kranke Zelle eine Onko-Zelle sein, insbesondere eine Onko-Zelle mit ALL (Akute lymphoblastische Leukämie).

Ferner kann auch die kranke Zelle ein Onko-Gen, insbesondere ein ALL-Onko-Gen, aufweisen.

Auch kann für eine Vielzahl von ausgewählten Genen des regulatorischen genetischen Netzwerks jeweils eine Gen-Expressionsrate vorgegeben werden, eine Vielzahl von resultierenden Gen-Expressionsmustern generiert werden und/oder eine Vielzahl von Vergleichen durchgeführt werden.

Bei einer Weiterbildung wird die Generierung der Vielzahl von resultierenden Gen-Expressionsmustern iterativ durchgeführt.

Ferner eignet sich die erfinderische Vorgehensweise oder Weiterbildung davon insbesondere zur Identifizierung eines dominanten Gens und/oder eines degenerierten/mutierten/kranken/onkogenen/Tumor-suppressor Gens.

Auch eignet sie sich zur Identifizierung einer Tumorzelle, beispielsweise im Zusammenhang mit einer Krebserkennung.

Ferner ist die erfinderische Vorgehensweise insbesondere geeignet zu einer Ursachenanalyse für ein abnormales Gen-Expressionsmuster/Gen-Expressrate.

Auch kann sie eingesetzt werden zu einer Simulation und/oder Analyse einer Wirkweise eines Medikaments.

In Figuren ist ein Ausführungsbeispiel der Erfindung dargestellt, welches im weiteren näher erläutert wird.

Es zeigen

Figur 1 eine Skizze einer Vorgehensweise bei der Untersuchung genetisch bedingter Krankheitsursachen durch Bayesssche inverse Modellierung am Beispiel einer Krebserkrankung;

Figur 2 eine Skizze mit einem Algorithmus zur Erzeugung eines Datensatzes von N Stichproben gemäß einem Ausführungsbeispiel;

Figur 3 eine Skizze für eine Vorgehensweise zur Erzeugung von Datensätze, welche eine Auswirkung von verschiedenen Beobachtungen widerspiegeln gemäß einem Ausführungsbeispiel;

Figuren 4a und b Skizzen die zeigen, dass durch eine Stichprobenentnahme gewonnene Daten Untertyp-charakteristische Expressionsmuster zeigen wie auch in einem ursprünglichen Datensatz;

Figur 5 eine Skizze, die graphisch zeigt eine Wahrscheinlichkeit jedes Untertyps unter einer Bedingung, dass ein Gen überexprimiert ist, für alle 271 Gene;

Figur 6 eine Skizze einer Graphenstruktur eines kausalen Netzwerks, welches ein regulatorisches genetisches Netzwerk repräsentiert.

Ausführungsbeispiel: Untersuchung genetisch bedingter Krankheitsursachen durch Bayessche inverse Modellierung am Beispiel einer Krebserkrankung (insb. Fig.1)

Überblick über die Vorgehendweise - Bayessche inverse Modellierung (BIM)

Auf vielen Gebieten der empirischen Forschung möchte man aus der Beobachtung von Versuchsergebnissen auf das zugrundeliegende Prinzip und dessen Ursprung schließen - die Beziehung zwischen "Ursache" und "Wirkung".

Zum Beispiel wird in der Krebsforschung das zugrundeliegende Prinzip studiert, welches bewirkt, dass sich eine normale Zelle in eine bösartige, schnell wachsende Krebszelle verwandelt.

Die Auswirkung der verschiedenen Arten des Krebses ist bekannt, z. B. das allgemeine Erscheinungsbild einer Krebszelle im Vergleich zu einer normalen Zelle, gemessen mit Hilfe von Microarray-Chips.

Dagegen ist die Ursache ihrer Entstehung größtenteils unbekannt.

Aufgrund der Einsicht, dass Krebs eine genetische Krankheit ist und dass er auf eine Abweichung des Verhaltens der Zellen zurückzuführen ist, konzentriert sich die Forschung auf die Aufdeckung der genetischen Prinzipien, die für die Entwicklung des Krebses verantwortlich sind.

Eine wichtige Aufgabe in diesem Umfeld ist es, Gene zu identifizieren, welche bei der Tumorgenese eine Rolle spielen können, wie etwa Geschwülste und Tumore unterdrückende Gene.

Nachfolgend wird eine Vorgehensweise beschrieben, mit der es möglich ist, Gene zu identifizieren, die eine potenzielle Ursache für die Tumorgenese sind.

Ein Element der Vorgehensweise ist ein statistisches Verfahren, in diesem Fall ein Bayesianisches (Bayessches) Netzwerk [3] (siehe obige und nachfolgende Ausführungen dazu), welches aus einem Microarray-Datensatz [1] gelernt wird [2] (siehe nachfolgend dazu "Strukturelles Lernen") (vgl. **Fig.1**).

Dabei wird angenommen, dass die Menge der gemessenen Genexpressionsvektoren X einer Grundgesamtheit mit einer hochdimensionalen multivariaten Wahrscheinlichkeitsdichtefunktion angehört, welche mit Hilfe eines Bayesschen Netzwerkes mit adaptiver Netzwerkstruktur modelliert wird.

Die Zusammenhänge zwischen den Variablen, nämlich die bedingten Abhängigkeiten und Unabhängigkeiten, werden mittels eines gerichteten azyklischen Graphen (directed acyclic graph, DAG) *G* dargestellt.

Die probabilistische Semantik eines Bayesschen Netzwerkes ist zur Analyse von Microarray-Daten sehr gut geeignet, da sie an die stochastische Natur sowohl der biologischen Prozesse als auch der mit einem Rauschen behafteten Experimente angepasst ist.

Bei der nachfolgend beschriebenen Vorgehensweise wird das gelernte Bayessche Netzwerk als ein generatives Modell zur Stichprobenentnahme von künstlichen Microarray-Datensätzen verwenden, welches die Dichteschätzung der gelernten bedingten Wahrscheinlichkeitsverteilungen liefert (vgl. **Fig.1, Schritte 110 - 130**).

Weiter wird der Effekt des Expressionszustandes bestimmter Gene auf das globale Expressionsmuster (inverse Modellierung) geschätzt, indem ein resultierende Datensatz analysiert wird (vgl. **Fig.1 Schritte 110 - 130**).

Auch wird bei der nachfolgend beschriebenen Vorgehensweise jedem Gen seine Wahrscheinlichkeit zugeordnet, mit der es die Ursache eines dieser Zellzustände ist.

Dazu werden diese Datensätze mit aus Microarray-Untersuchungen von verschiedenen bekannten Zellzuständen erhaltenen Daten verglichen (vgl. **Fig.1, Schritt 130**).

Anschaulich gesehen, konzentriert sich die Vorgehensweise nicht explizit auf die Struktur des Netzwerkes, sondern vielmehr auf die Wahrscheinlichkeitsverteilung, die durch das gelernte Bayessche Netzwerk abgeleitet wird.

Schließlich wird die Vorgehensweise auf Microarray-Daten von verschiedenen Untertypen von pädiatrischer akuter Lymphoblasten-Leukämie (ALL) von Yeoh et al. [4] angewendet.

Durch den Vergleich der künstlichen Daten mit Expressionsmustern von spezifischen Krebs-Untertypen erhält man ein Wahrscheinlichkeitsmaß des krankheitserzeugenden Verhaltens jedes Gens (vgl. **Fig.1, Schritt 130**).

Ergebnisse der angewendeten Vorgehensweise zeigen, dass diese in Verbindung mit der Bayesschen inversen Modellierung (BIM) es ermöglicht, die Auswirkung von pathogenetisch veränderten Expressionsniveaus auf das globale Expressionsmuster vorherzusagen, wobei bereits bekannte Onkogene ebenso wie potenziell neue gefunden werden.

Bayessche Netzwerke

Im Obigen wurden bereits Grundlagen von Bayesschen Netzen [3] beschrieben.

Im Falle der Modellierung eines regulierenden genetischen Netzwerkes durch ein Bayessches Netzwerk werden Gene bzw. ihre entsprechenden Proteine durch Knoten symbolisiert.

Regelungsmechanismen werden durch Kanten zwischen zwei Knoten beschrieben, welche auf eine kausale Art und Weise interpretiert werden können.

Die Qualität der Regulierung ist in der bedingten Wahrscheinlichkeitsverteilung des betroffenen Gens bei gegebenen Regulatoren desselben codiert.

Strukturelles Lernen

Der Vorgang des strukturellen Lernens kann wie folgt beschrieben werden:

Sei $D = \{d^1, d^2, \ldots, d^N\}$ ein Datensatz von N unabhängigen Beobachtungen, wobei jeder Datenpunkt ein n-dimensionaler Vektor mit Komponenten $d^1 = \{d^1_1, d^1_2, \ldots, d^1_n\}$ ist. Bei gegebenem D ist die Struktur G des Bayesschen Netzwerkes zu finden, welche am besten mit D übereinstimmt, d. h. welche die Bayes-Punktbewertung (Bayes-Score)

$$S(G|D) = \frac{P(D|G)P(G)}{P(D)} \qquad (2)$$

maximiert, wobei $P(D|G)$ die Randwahrscheinlichkeit, $P(G)$ die Apriori-Wahrscheinlichkeit der Struktur und $P(D)$ die Evidenz ist.

Da sowohl die Apriori-Wahrscheinlichkeit als auch die Evidenz unbekannt sind, reduziert sich das Problem auf das Ermitteln der Struktur mit der besten Randwahrscheinlichkeit entsprechend den Daten (Heckerman et al. [5]).

Wenn der Datensatz D aus N Microarray-Experimenten besteht, z. B. aus Zellproben von unterschiedlichen Patienten, repräsentiert jeder Datenvektor $\{d^1_1, d^1_2, \ldots, d^1_n\}$ das Expressionsprofil von n Genen in einem Microarray-Experiment.

Ein aus solchen Daten gelerntes Bayessches Netzwerk codiert die Wahrscheinlichkeitsverteilung von n Genen, die aus diesen N Microarray-Experimenten erhalten wurden.

Bayessche inverse Modellierung (BIM)

Generatives Modell

Ein gelerntes (siehe obige Ausführungen zu "Strukturelles Lernen") Bayessches Netzwerk B stellt eine Dichte-Schätzfunktion dar, welche die Wahrscheinlichkeitsverteilung des Datensatzes D, von dem ausgehend es gelernt wurde, mit Hilfe der Menge der bedingten WDFen widerspiegelt.

Somit kann es als ein generatives Modell zur Erzeugung eines Datensatzes D_B verwendet werden, welcher die aus D erhaltene Dichteverteilung widerspiegelt.

Fig.2 zeigt einen Algorithmus 200 zur Erzeugung eines Datensatzes von N Stichproben aus B.

Der erste Schritt 210 des Algorithmus 200 besteht darin, alle Variablen so zu ordnen, dass die Parents (Elternknoten) Pa_i vor X_i instantiiert werden.

Anschließend werden die Variablen entsprechend der Ordnung gewählt und mit einem Wert instantiiert 220.

Der Wert jeder Variablen wird mit Wahrscheinlichkeit $P(\text{Zustand}|Pa_i)$ gewählt. Dieser Schritt wird wiederholt 230 , bis N Stichproben erzeugt sind.

Probabilistische Interferenz

Ein wesentliches Problem in Bayesschen Netzwerken ist die Evidenz-Fortpflanzung, das heißt, die Ermittlung der Aposteriori-Verteilungen $P(X_q|E)$ einer Abfragevariablen X_q, wenn eine gewisse Evidenz E im Bayesschen Netzwerk beobachtet worden ist.

Aufgrund der Definition einer bedingten Wahrscheinlichkeit ist die Aposteriori-Wahrscheinlichkeit

(3) $$P(X_q|E) = \frac{P(X_q, E)}{P(E)} = \frac{\sum_{X\setminus\{X_q, X_E\}} P(X)}{\sum_{X\setminus X_E} P(X)}$$

wobei X_E die Menge der beobachteten Variablen bezeichnet.

Um die Zeitkomplexität zu überwinden, verwenden die verschiedenen Methoden der exakten Interferenzberechnung das allgemeine Prinzip der dynamischen Programmierung.

Im Rahmen dieses Ausführungsbeispiel wird ein einfacher Interferenzalgorithmus, der "bucket elimination" [6], verwendet.

Die Grundidee bei diesem Interferenzalgorithmus besteht darin, Variablen eine nach der anderen entsprechend einer Eliminationsreihenfolge ρ durch Summieren zu eliminieren.

Auf diese Weise kann $P(X_q|E)$ innerhalb einer annehmbaren Zeit effizient berechnet werden.

Interventionelle Modellierung durch Einstellen der Evidenz

Bei der Herangehensweise der interventionellen Modellierung wird die Auswirkung einer bestimmten Beobachtung auf das Verhalten des Bayesschen Netzwerkes unter Verwendung einer Kombination von probabilistischer Interferenz und Daten-Stichprobenentnahme geschätzt.

Entsprechend **Fig.3** kann das Bayessche Netzwerk als eine Art Black Box 300 angesehen werden, wobei der Eingang durch eine Menge von Beobachtungen E 310 und die entsprechende Liste von beobachteten Variablen X_E 320 gegeben ist.

Der Ausgang, der durch den Datensatz $D_{B|E}$ 330 gegeben ist, wird wie im Vorigen zugehörig zu **Fig.2** beschrieben erzeugt.

Zusätzlich ist die beobachtete Evidenz zu berücksichtigen.

Folglich wird jeder Zustand von X_i mit Wahrscheinlichkeit P(Zustand|Pa_i,E) gewählt, welche mittels probabilistischer Interferenz berechnet wird.

Mit beschriebener Vorgehensweise gemäß **Fig.3** können nun unterschiedliche Datensätze erzeugt werden, welche die Auswirkung der verschiedenen Beobachtungen widerspiegeln.

Wenn wie nachfolgend beschrieben biologische Auswirkungen analysiert werden, heißt das, dass durch diese Vorgehensweise gemäß **Fig.3** künstliche Microarray-Daten erzeugbar sind, welche die Wahrscheinlichkeitsverteilung eines gewissen Datensatzes widerspiegeln, wenn bestimmte Beobachtungen gegeben sind.

Vergleicht man die künstlich erzeugen Daten mit Daten von bekannter Herkunft, z. B. mit einer krebsspezifischen Menge von Messdaten, können jene Gene bestimmt werden, welche, wenn sie auf einem gewissen Expressionsniveau fixiert werden, das Modell so beeinflussen, dass die beiden Microarray-Datensätze, der künstliche und der bekannte, dieselben Eigenschaften aufweisen.

Statistischer Vergleich von Datensätzen

Um die Qualität des Einflusses der Evidenz E auf das Verhalten des Bayesschen Netzwerkes B zu schätzen, wird der erzeugte Datensatz $D_{B|E}$ mit einer Menge von Datensätzen D von bekannten Zuständen S verglichen.

Es wird angenommen, dass D die Auswirkung verschiedener Krebsarten beschreibt. Ausführungsgemäß kann nun das Verhalten von Evidenz E in Bezug auf eine bestimmte Krebsart S beschrieben werden.

Unter Verwendung eines Abstandsmaßes wird die Änderung a der Korrelation zwischen $D_{B|E}$ und D_S infolge von E schätzbar:

(4) $$a(E) = \frac{d(D_{B|E}, D_S)}{d(D_B, D_S)}$$

wobei der Abstand zwischen den zwei Datensätzen mit Hilfe des Abstands zwischen D_B, welches aus B ohne Evidenz entnommen wurde, und D_S normiert wurde.

Folglich ist ausführungsgemäß der Einfluss einer beobachteten Evidenz messbar, z. B. der Expressionszustand eines bestimmten Gens auf ein für Krebs charakteristisches Verhalten des Modells.

Zweitens ist die Wahrscheinlichkeit dafür berechenbar, dass B einen Datensatz $D_{B|E}$ erzeugt, welcher gleich D_S bei gegebenem E ist.

Zu diesem Zweck wird geschätzt, wie viele Stichproben d^1 von $D_{B|E}$ am nächsten bei D_S liegen, indem der Abstand zwischen jeder Stichprobe und jedem Datensatz von D berechnet wird.

Somit erhält man die Aposteriori-Wahrscheinlichkeit $P(S|E)$ des Auftretens der Krebsart S bei gegebener Evidenz E aus:

(5) $$P(S|E) = \frac{N_{ES}}{N}$$

wobei N_{ES} die Anzahl der Stichproben von $D_{B|E}$ ist, welche statistisch dem Datensatz D_S am nächsten kommen, und wobei N die Gesamtzahl der Stichproben von $D_{B|E}$ ist.

Wie bereits im Obigen konstatiert beschäftigt sich die empirische Forschung mit der Beziehung zwischen Ursache und Wirkung, indem sie aus einer experimentellen Beobachtung Rückschlüsse auf die zugrundeliegende Ursache zieht.

Mit der Herangehensweise der Bayesschen inversen Modellierung gemäß dem Ausführungsbeispiel wird eine zugrundeliegende Ur-

sache geschätzt, indem zuerst eine Wirkung erzeugt wird, die aus einer bekannten Beobachtung hervorgeht.

Nach diesem inversen Schritt wird diese Wirkung mit Wirkungen verglichen, welche wohldefiniert sind, deren Ursache jedoch unbekannt ist.

Die potenzielle Ursache der am besten übereinstimmenden Wirkung ist dann durch die Beobachtung gegeben, welche die erzeugte Wirkung hervorruft.

Der ALL-Microarray-Datensatz von Yeoh et al. [4]

Die Daten, die für die Analyse gemäß dem Ausführungsbeispiel verwendet werden, bestehen aus 327 Stichproben von verschiedenen Untertypen von pädiatrischer akuter Lymphoblasten-Leukämie (ALL).

Der Datensatz wurde von Yeoh und seinen Kollegen vom St. Jude Children's Research Hospital [4] zusammengestellt.

ALL ist eine heterogene Krankheit, die verschiedene Untertypen umfasst, einschließlich sowohl Leukämie vom T-Zelltyp als auch Leukämie vom B-Zelltyp, die sich hinsichtlich ihrer Reaktion auf eine medizinische Behandlung deutlich unterscheiden.

Abgesehen von T-ALL, deren Ursache noch nicht klar bekannt ist, kann jeder B-Zellen-Untertyp auf eine spezifische genetische Veränderung zurückgeführt werden, z. B. auf genetische Translokationen t(9;22) [BCR-ABL], t(1;19) [E2A-PBX1], t(12;21) [TEL-AML1], t(4;11) [MLL] oder auf einen hyperdiploiden Karyotyp [> 50 Chromosomen].

Daher ist es nicht verwunderlich, dass die Expressionsmuster der verschiedenen Untertypen recht deutlich voneinander unterscheiden.

Ferner zeigen Microarray-Daten noch ein anderes deutliches Expressionsprofil, welches auf die Existenz eines weiteren ALL-Untertyps zusätzlich zu den 6 bekannten hindeutet.

Es soll angemerkt werden, dass Yeoh et al. [4] an einem robusten Klassifikator zur Klassifizierung der Untertypen unter Verwendung einer Stützvektor-Maschine mit einem Satz von 271 diskriminierenden Genen arbeitet.

Ergebnisse

Gelernte Struktur

Für die Analyse gemäß dem Ausführungsbeispiel wird der reduzierte Datensatz von 271 Genen und 327 Stichproben von verschiedenen ALL-Untertypen [4], wie oben beschrieben, verwendet.

Um den Lernvorgang eines multivariaten Modells durchzuführen, wurde der Datensatz in die Werte "unterexprimiert", "normal exprimiert" und "überexprimiert" diskretisiert.

Die gelernte Struktur zeigt "maßstabfreie" (scale-free) Kenngrößen, ein Merkmal, welches für biologische Netze, wie etwa für metabolische Netze oder Signalisierungsnetze, typisch ist.

Solche Netze sind durch eine Potenzverteilung des Grades (Ranges) eines Knotens gekennzeichnet, welcher als die Anzahl der Verbindungen mit anderen Knoten definiert ist.

Diese Knoten besitzen einen starken Einfluss auf die Dynamik und Robustheit von "maßstabfreien" Netzen, und von vielen dieser in starkem Maße verbundenen Gene in unserem Modell ist tatsächlich bekannt, dass sie eine Rolle bei der Onkogenese oder bei mit der Krebsentwicklung zusammenhängenden kritischen Prozessen spielen, z. B. DNA-Reparatur.

Zuerst wird nun ein Datensatz von 300 Stichproben aus dem Modell erzeugt, um die Statistiken zu schätzen, die durch die Menge der bedingten Wahrscheinlichkeiten definiert sind.

Fig.4 zeigt, dass die durch die Stichprobenentnahme gewonnenen Daten (**Fig.4b**) Untertyp-charakteristische Expressionsmuster zeigen, so wie dies auch im ursprünglichen Datensatz (**Fig.4a**) der Fall ist.

Die Muster einiger Untertypen, wie etwa E2A-PBX1 oder T-ALL, werden sehr gut reproduziert, während einige andere weniger gut generiert werden, z. B. das Muster des Untertyps MLL, oder völlig verfehlt werden, wie etwa BCR-ABL.

Modellierung von Leukämie-Untertypen durch Intervention

Das gelernte Bayessche Netzwerk ist die Ausgangsbasis bei dem Ausführungsbeispiel für die Herangehensweise, mittels inverser Modellierung diejenigen Gene zu finden, welche, wenn sie

auf einem bestimmten Expressionsniveau fixiert werden, das Modell so beeinflussen, dass der generierte künstliche Microarray-Datensatz spezifische Merkmale aufweist.

Wie im Obigen beschrieben wurde, wird die Wahrscheinlichkeit $P(C|E)$ der Erzeugung eines bestimmten Krebs-Untertyps C geschätzt, wenn eine gewisse Beobachtung E gegeben ist, in diesem Falle der Expressionszustand eines bestimmten Gens $P(C|Gen_i=Zustand)$.

Im Gegensatz zu Yeoh wird nicht nur das Vorliegen eines bestimmten Krebs-Untertyps vorhergesagt, sondern genetische Mechanismen, die zu seiner Erzeugung führen.

Eine hohe Wahrscheinlichkeit sagt voraus, dass die fixierten Gene eine potenzielle Ursache für das Untertyp-spezifische Expressionsverhalten der fraglichen Gene ist, welches wiederum die zugrundeliegende Ursache für ein spezifisches kanzeröses Erscheinungsbild sein kann.

Für den Vergleich werden 7 Referenz-Datensätze verwendet, wobei jeder von ihnen in Verbindung mit einem spezifischen ALL-Untertyp erhalten wurde.

Fig.4a zeigt, dass der ursprüngliche Microarray-Datensatz deutlich in 7 Cluster (Punkthaufen) mit unterschiedlichen Stichprobenumfängen unterteilt ist.

Jeder dieser Cluster repräsentiert das Expressionsmuster von 271 Genen, wenn ein bestimmter Leukämie-Untertyp gegeben ist, und wurde verwendet, um den Einfluss einer Evidenz auf das Auftreten dieser verschiedenen ALL-Untertypen zu messen.

In einem ersten Schritt wird jedes Gen bei irgendeinem seiner Expressionswerte fixiert, wobei alle diese Bedingungen verwendeten werden, um einen Datensatz von 300 Stichproben zu generieren (**Fig.4b**).

Anschließend werden alle diese Daten mit den 7 Referenz-Datensätzen, wie vormals erläutert, verglichen.

In **Fig.5** ist die Wahrscheinlichkeit jedes Untertyps unter der Bedingung, dass ein Gen überexprimiert ist, für alle 271 Gene graphisch dargestellt.

Fig.5 zeigt, dass eine kleine Anzahl von Genen existiert, welche einen bestimmten ALL-Untertyp mit einer hohen Wahrscheinlichkeit hervorrufen, wenn sie stark aktiv sind.

Um diese Ergebnisse zu beweisen, wird die molekulare Funktion gewisser Gene und ihre Rolle in biologischen Prozessen, insbesondere im Hinblick auf die Pathogenese, nachfolgend eingehender betrachtet.

Biologische Einblicke

Dazu werden die Gene näher betrachtet, die mit einer hohen Wahrscheinlichkeit einen bestimmten Untertyp verursachen, sowie signifikante Strukturmuster in dem gelernten Netzwerk, d. h. dominante Gene und ihre Umgebung.

Das gelernte Bayessche Netzwerk (Modell) resultiert aus einem Microarray-Datensatz von unterschiedlichen Leukämie-Untertypen und spiegelt transskriptionale Beziehungen zwischen Genen wider, die in diesen bösartigen Krebszellen auftreten.

Somit sind Gene, die einen bestimmten Untertyp hervorrufen, entweder potenzielle Onkogene oder werden durch solche Gene reguliert.

Das erste Gen, welches eingehender analysiert wird, ist das Gen PBX1.

Wenn es überexprimiert ist, erzeugt das gelernte Bayessche Netzwerk mit einer Wahrscheinlichkeit von 0,96 einen Datensatz, welcher für den Untertyp E2A-PBX1 der ALL vom B-Zelltyp charakteristisch ist (siehe **Fig.5**).

Dies legt die Vermutung nahe, dass ein kausaler Zusammenhang zwischen der "Überexprimiertheit" dieses Gens und dem Auftreten des ALL-Untertyps E2A-PBX1 vorhanden ist.

Und tatsächlich ist PBX1 als ein Protoonkogen bekannt, welches die Verwandlung von normalen Blutzellen in bösartige ALL-Krebszellen verursacht.

Infolge der Chromosomen-Translokation t(1;19) verschmilzt PBX1 mit dem Gen E2A und verwandelt sich in ein potentes Onkogen, welches den Leukämie-Untertyp E2A-PBX1 verursacht.

Da ferner die Graphstruktur des Modells (**Fig.6**) auf eine kausale Weise interpretiert werden kann, liefert sie Informationen über die Wechselwirkung zwischen potenziellen Onkogenen und anderen Genen, was wiederum als eine onkogene Regelung interpretiert werden kann.

Wenn man die Struktur des Netzwerkes (**Fig.6**) betrachtet, so stellt PBX1 ein dominantes Gen dar, indem es viele andere Ge-

ne beeinflusst, jedoch nur von einem oder wenigen anderen Genen reguliert wird.

Zusätzlich identifiziert das Modell aufgrund der bedingten Wahrscheinlichkeitsverteilung PBX1 als einen Transkriptionsaktivator.

Dies kann ebenfalls durch bekannte biologische Tatsachen erklärt werden, da PBX1 Gene aktiviert, die normalerweise entweder nicht exprimiert oder auf einem niedrigen Niveau exprimiert sind.

Patienten mit einer Hyperdiploidie von > 50 Chromosomen haben Klone von 51-68 Chromosomen. Obwohl hoch hyperdiploide Klone selten identisch sind, neigen sie dazu, ein Muster des Chromosomenzuwachses mit zusätzlichen Kopien der Chromosome 4, 6, 10, 14, 18 und 21 aufzuweisen.

Trisomie und Polysomie 21 sind nicht zufällige Anomalien, welche bei ALL häufig zu beobachten sind. Ihr Auftreten, auch wenn es nicht spezifisch ist, sowie das gehäufte Auftreten von akuter Leukämie bei Subjekten mit konstitutioneller Trisomie 21 legen die Vermutung nahe, dass das Chromosom 21 eine besondere Rolle bei der Leukämogenese spielt.

Eine andere Krankheit, das Down-Syndrom, wird durch Trisomie 21 verursacht und zeigt ein verstärktes Auftreten von Leukämie wie etwa ALL.

Demzufolge ermöglich in diesem Fall die beschriebene Vorgehensweise gemäß dem Ausführungsbeispiel Gene zu identifizieren, die in hohem Maße auf den hyperdiploiden ALL-Untertyp hinweisen, von denen jedoch auch bekannt ist, dass sie eine

wesentliche Rolle bei der Entstehung des Down-Syndroms spielen.

Das Gen SOD1 befindet sich am Chromosom 21 und produziert ein Enzym, welches superoxidfreie Radikale in Wasserstoffperoxid umwandelt. Die verstärkte Expression bei Trisomie 21, welche auch bei den Microarray-Stichproben von Patienten mit hyperdiploidem Karyotyp zu beobachten ist, kann die Hirnschädigung auslösen, die beim Down-Syndrom zu erkennen ist.

Die Häufigkeit des Auftretens des hyperdiploiden ALL-Untertyps erhöht sich auch in dem Falle, wenn das Gen PSMD10 überexprimiert ist.

PSMD10 ist eine regulierende Unter-Einheit des Proteasoms 26S, von dem nachgewiesen wurde, dass es als ein natürlicher Mechanismus für den Abbau von Proteinen durch Regulierung des Proteinumsatzes in eukaryotischen Zellen wirkt.

Dies ist bei Krebserkrankungen des Menschen von Bedeutung, da der Zellzyklus, das Tumorwachstum und das Überleben durch eine große Vielfalt an intrazellulären Proteinen bestimmt werden, welche durch den Ubiquitin-abhängigen Proteasom-Abbauweg geregelt werden, der von PSMD10 beeinflusst wird.

In neueren wissenschaftlichen Arbeiten auf diesem Gebiet wurde nachgewiesen, dass dieser Abbauweg oft Gegenstand einer mit Krebs zusammenhängenden Deregulierung ist und solchen Prozessen unterliegen kann, wie onkogener Transformation, Tumorprogression, Umgehung der Immunüberwachung und Arzneimittelresistenz.

Zusammenfassung des Ausführungsbeispiels

Das beschriebene Ausführungsbeispiel stellt eine neue Vorgehensweise vor, mit der es möglich ist, Gene, die eine potenzielle Ursache für eine Tumorgenese sind, durch Analysieren der Zusammenhänge zwischen Microarray-Daten von Leukämie-Untertypen und einem Datensatz, der Ergebnis einer Stichprobenentnahme aus einem gelernten Bayesschen Netzwerk ist, zu identifizieren.

Basis dieser Vorgehensweise ist die Modellierung eines regulierenden genetischen Netzwerkes durch ein Bayessches Netzwerk, wobei Gene bzw. ihre entsprechenden Proteine durch Knoten des Bayesschen Netzwerks symbolisiert werden.

Regelungsmechanismen werden durch Kanten zwischen zwei Knoten beschrieben, welche auf eine kausale Art und Weise interpretiert werden können.

Die Qualität der Regulierung ist in der bedingten Wahrscheinlichkeitsverteilung des betroffenen Gens bei gegebenen Regulatoren desselben codiert.

Das Verständnis der regulierenden genetischen Netze stellt einen wichtigen Schritt auf dem Weg zur Charakterisierung der genetischen Mechanismen dar, welche komplexen Krankheiten zugrunde liegen.

In der Krebsforschung, wo die Identifizierung von Geschwülste und Tumore unterdrückenden Genen einen Schlüsselrolle spielt, ist die Kenntnis neuer potenzieller Onkogene und ihrer Wechselwirkung mit anderen Molekülen ein wichtiger Beitrag zur Aufdeckung der Grundprinzipien, welche die Umwandlung normaler Zellen in bösartige Krebszellen bestimmen.

Mit der beschriebene Vorgehensweise gemäß dem Ausführungsbeispiel, insbesondere mit der Bayesschen inversen Modellierung, ist es möglich, Gene mit einer solchen onkogenen Charakteristik einfach durch eine statistische Analyse von Gen-Expressionsmustern, die mit Hilfe von DNA-Microarrays gemessen wurden, zu entdecken.

Das zugrundliegende wahrscheinlichkeitstheoretische Modell, das verwendet wurde, ist ein Bayessches Netzwerk, welches die multivariate Wahrscheinlichkeitsverteilung einer Menge von Variablen mittels einer Menge von bedingten Wahrscheinlichkeitsverteilungen codiert.

Die statistischen Abhängigkeiten werden in einer Graphstruktur codiert. Beim Lernverfahren werden Bayessche Statistiken verwendet, um die Netzstruktur und die entsprechenden Modellparameter zu ermitteln, welche die Wahrscheinlichkeitsverteilung enthalten in den Daten am besten beschreiben.

In diesem Dokument sind folgende Schriften zitiert:

[1] Stetter Martin et al., Large-Scale Computational Modeling of Generic Regulatory Networks, Kluwer Academic Publisher, Niederlande, 2003;

[2] Offenlegungsnummer DE 10159262.0;

[3] F. W. Jensen, F. V. (1996), An introduction to Bayesian networks, UCL Press, London; 178 pages;

[4] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Petal et al. (2002), Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell 1:133-143;

[5] D. Heckerman, D. Geiger and D. Chickering (1995), Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20:197-243;

[6] R. Dechter (1996), Bucket elimination: A unifying framework for probabilistic inference. In: Uncertainty in Artificial intelligence, UA196:211-219.

Patentansprüche

1. Verfahren zur Analyse eines regulatorischen genetischen Netzwerks einer Zelle unter Verwendung eines kausalen Netzes, welches kausale Netz das regulatorische genetische Netzwerk der Zelle beschreibt derart, dass Knoten des kausalen Netzes Gene des regulatorischen genetischen Netzwerks repräsentieren und Kanten des kausalen Netzes regulatorische Wechselwirkungen zwischen den Genen des regulatorischen genetischen Netzwerks repräsentieren,

a) bei dem für ein ausgewähltes Gen des regulatorischen genetischen Netzwerks eine Gen-Expressionsrate vorgegeben wird,

b) bei dem unter Verwendung des kausalen Netzes für die vorgegebene Gen-Expressionsrate ein resultierendes Gen-Expressionsmuster für das regulatorische genetische Netzwerk generiert wird;

c) bei dem das generierte resultierende Gen-Expressionsmuster mit einem vorgegebenen Gen-Expressionsmuster des regulatorischen genetischen Netzwerks verglichen wird.

2. Verfahren nach Anspruch 1,
bei dem das ausgewählte Gen unter Verwendung des kausalen Netzes mittels einer Abhängigkeitsanalyse ausgewählt wird.

3. Verfahren nach einem der vorangehenden Ansprüche,
bei dem die Gen-Expressionsrate des ausgewählten Genes derart vorgegeben wird, dass die vorgegebene Gen-Expressrate des ausgewählten Genes eine Annahme eines Gendefekts widerspiegelt.

4. Verfahren nach einem der vorangehenden Ansprüche,
bei dem das kausales Netz ein Bayesianisches Netz ist.

5.Verfahren nach einem der vorangehenden Ansprüche,

bei dem das kausale Netz von einem Typ DAG (directed acylic graph) ist.

6. Verfahren nach einem der vorangehenden Ansprüche, bei dem das generierte resultierende und/oder das vorgegebene Gen-Expressionsmuster diskrete Genzustände repräsentiert.

7. Verfahren nach einem der vorangehenden Ansprüche, -bei dem die repräsentierten diskreten Genzustände ein über-, ein normal-, ein unterexprimierten Genzustand sind

8. Verfahren nach einem der vorangehenden Ansprüche, bei dem der Vergleich des generierten resultierenden Gen-Expressionsmuster mit dem vorgegebenen Gen-Expressionsmuster unter Verwendung eines statischen Verfahrens und/oder einer statistischen Kennzahl, insbesondere eines Abstandsmaßes, durchgeführt wird.

9. Verfahren nach einem der vorangehenden Ansprüche, bei dem das kausales Netz unter Verwendung von Gen-Expressionsmustern trainiert wird, wobei die Knoten und die Kanten des kausalen Netzes angepasst werden.

10. Verfahren nach einem der vorangehenden Ansprüche, bei dem die Gen-Expressionsmuster, insbesondere das vorgegebene Gen-Expressionsmuster und/oder die Gen-Expressionsmuster für das Training, bestimmt werden unter Verwendung einer DNA-Micro-Array-Technik.

11. Verfahren nach einem der vorangehenden Ansprüche, bei dem das vorgegebene Gen-Expressionsmuster und/oder die Gen-Expressionsmuster für das Training Gen-Expressionsmuster eines genetischen regulatorischen Netzwerks einer kranken Zelle ist.

12. Verfahren nach einem der vorangehenden Ansprüche,

bei dem die kranke Zelle eine Onko-Zelle, insbesondere eine Onko-Zelle mit ALL (Akute lymphoblastische Leukämie) ist.

13. Verfahren nach einem der vorangehenden Ansprüche, bei dem die kranke Zelle ein Onko-Gen, insbesondere ein ALL-Onko-Gen, aufweist.

14. Verfahren nach einem der vorangehenden Ansprüche, bei dem für eine Vielzahl von ausgewählten Genen des regulatorischen genetischen Netzwerks jeweils eine Gen-Expressionsrate vorgegeben wird, eine Vielzahl von resultierenden Gen-Expressionsmustern generiert werden und eine Vielzahl von Vergleichen durchgeführt werden.

15. Verfahren nach einem der vorangehenden Ansprüche, bei dem die Generierung der Vielzahl von resultierenden Gen-Expressionsmustern iterativ durchgeführt wird.

16. Verfahren nach einem der vorangehenden Ansprüche, eingesetzt zur Identifizierung eines dominanten Gens.

17. Verfahren nach einem der vorangehenden Ansprüche, eingesetzt zur Identifizierung eines degenerierten/mutierten/kranken/onkogenen/tumor-suppressor Gens.

18. Verfahren nach einem der vorangehenden Ansprüche, eingesetzt zur Identifizierung einer Tumorzelle.

19. Verfahren nach einem der vorangehenden Ansprüche, eingesetzt zur Krebserkennung.

20. Verfahren nach einem der vorangehenden Ansprüche, eingesetzt zu einer Ursachenanalyse für ein abnormales Gen-Expressionsmuster/Gen-Expressrate.

21. Verfahren nach einem der vorangehenden Ansprüche,

eingesetzt zu einer Simulation und/oder Analyse einer Wirkweise eines Medikaments.

22. Computerprogramm mit Programmcode-Mitteln, um alle Schritte gemäß Anspruch 1 durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

23. Computerprogramm mit Programmcode-Mitteln gemäß dem vorangehenden Anspruch, welche Programmcode-Mitteln auf einem computerlesbaren Datenträger gespeichert sind.

24. Computerprogramm-Produkt mit auf einem maschinenlesbaren Träger gespeicherten Programmcode-Mitteln, um alle Schritte gemäß Anspruch 1 durchzuführen, wenn das Programm auf einem Computer ausgeführt wird.

Zusammenfassung

Verfahren, Computerprogramm mit Programmcode-Mitteln und Computerprogramm-Produkt zur Analyse eines regulatorischen genetischen Netzwerks einer Zelle

Die Erfindung betrifft eine Analyse eines regulatorischen genetischen Netzwerks einer Zelle unter eines kausalen Netzes.

Bei dem Analyseverfahren wird für ein ausgewähltes Gen des regulatorischen genetischen Netzwerks eine Gen-Expressionsrate vorgegeben. Unter Verwendung des kausalen Netzes wird für die vorgegebene Gen-Expressionsrate ein resultierendes Gen-Expressionsmuster für das regulatorische genetische Netzwerk generiert. Das generierte resultierende Gen-Expressionsmuster wird anschließend mit einem vorgegebenen Gen-Expressionsmuster des regulatorischen genetischen Netzwerks verglichen.

Sign. Fig.1

200309564

Fig. 1

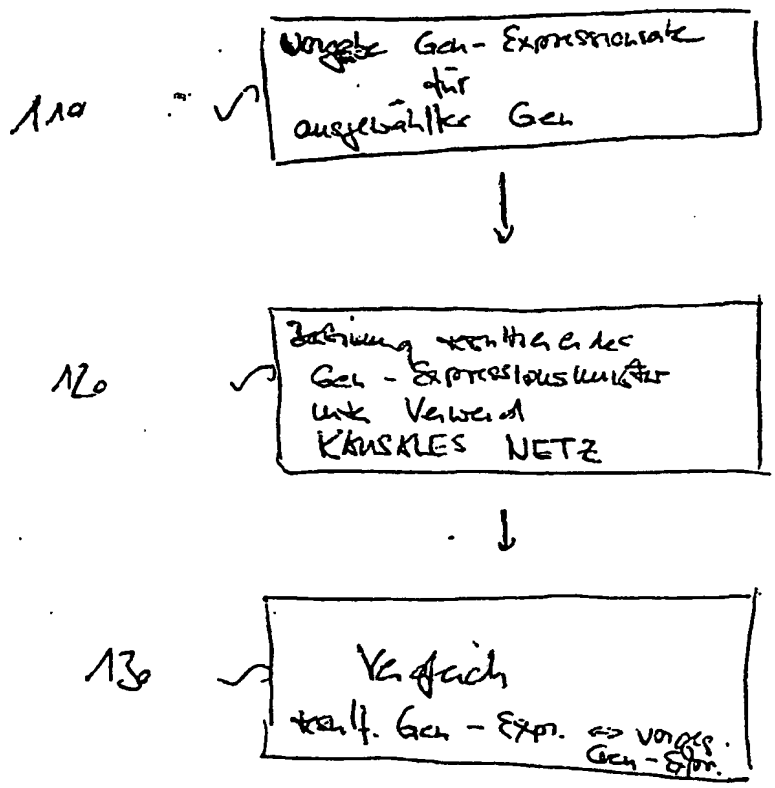


Fig.2

Algorithmus der Stichprobenentnahme (B,N)

Eingang:

B – Bayessches Netzwerk;

N – Anzahl der unabhängigen Stichproben.

Ausgang:

D_B - Datensatz von *N* unabhängigen Stichproben.

1. Ordne die Variablen-Menge X gemäß der Bedingung, dass Parents (Elternknoten) Pa_i vor den X_i angeordnet sind.
2. Für $s = 1, ..., N$
3. Für $i = 1, ..., n$
4. Sei X_i der Knoten mit der höchsten Ordnungsnummer in dieser Stichprobe, der nicht instantiiert ist.
5. Falls X_i ein Wurzelknoten ist, wähle den Zustand mit Wahrscheinlichkeit *P*(*Zustand*),
6. andernfalls wähle den Zustand mit Wahrscheinlichkeit *P*(Zustand|entnommene Zustände von Pa_i).
7. Instantiiere X_i = Zustand.

Fig.2: Algorithmus zur Erzeugung eines Datensatzes von *N* Stichproben aus einem Bayesschen Netzwerk *B*.

Fig.3

Algorithmus der interventionellen Stichprobenentnahme (B,E,N)

Eingang:

B – Bayessches Netzwerk;

E – Menge von Beobachtungen;

N – Anzahl der unabhängigen Stichproben.

Ausgang:

$D_{B|E}$ - Datensatz von *N* unabhängigen Stichproben bei gegebenem *E*.

X_E – Menge beobachteter Variabler;

$X_q = \{X \backslash X_E\}$ – Menge von Abfragevariablen.

1. Ordne X_q gemäß der Bedingung, dass Parents (Elternknoten) Pa_i vor den X_i angeordnet sind.
2. Für $s = 1, ..., N$
3. Für $i = 1, ..., n$
4. Sei X_i der Knoten mit der höchsten Ordnungsnummer in dieser Stichprobe, der nicht instantiiert ist.
5. Falls X_i ein Wurzelknoten ist, wähle den Zustand mit Wahrscheinlichkeit $P(Zustand|E)$,
6. andernfalls wähle den Zustand mit Wahrscheinlichkeit $P(\text{Zustand}|\text{entnommene Zustände von } Pa_i, E)$.
7. Instantiiere X_i = Zustand.

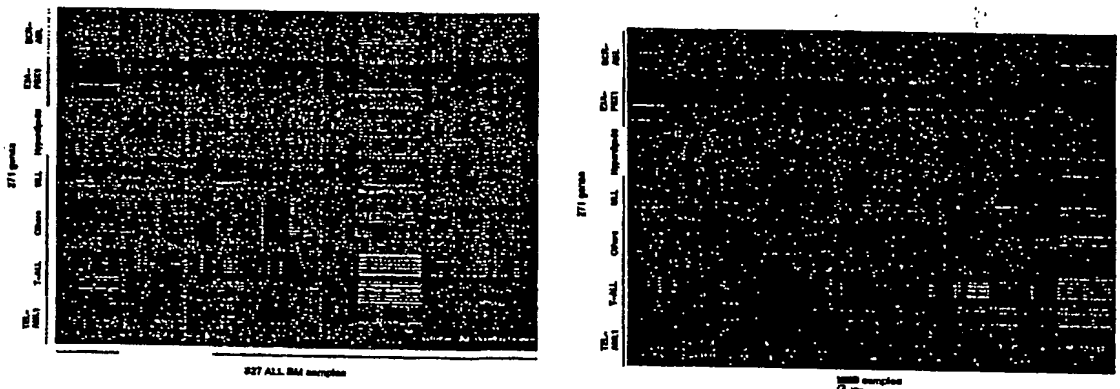
Fig.3.: Algorithmus zur Erzeugung eines Datensatzes von *N* Stichproben aus einem Bayesschen Netzwerk *B* bei gegebener Evidenz *E*.

3 Results

3.1 Learned structure

For our experiments we use the reduced dataset of 271 genes and 327 samples of different ALL-subtypes as described above. To train a multivariate model the dataset was discretized into the values over-, normal- and overexpressed.
The learned structure shows scale-free characteristics a feature that is typical for biological networks, such as metabolic or signaling networks. Such networks are characterized by a power-law distribution of a node's degree, defined as the number of connections with other ones. This nodes strongly affect the dynamics and robustness of scale-free networks and many of this highly connected genes in our model are in fact known to play a role in oncogenesis or in critical processes related to cancer development, e.g. DNA-repair.
First, we generate a dataset of 300 samples from the model to estimate the statitics defined by the set of conditional probabilites. In Figure 3 one can see that the sampled data shows subtype-characteristic expression-patterns as in the original one. The patterns of some subtypes, such as E2A-PBX1 or T-ALL, are strongly reproduced whereas some others are generated poorly, e.g. the pattern of subtype MLL, or completely missed such as BCR-ABL.

Fig. 4



a) original b) sampled

Figure 3: a: Original discretized ALL microarray-dataset showing the expression pattern of 271 genes over 372 bone marrow samples. b: Dataset of 300 samples generated from a learned network. Data shows subtype-characteristic expression-patterns.

3.2 Modeling leukemia subtypes by intervention

The learned Bayes-net is the basis for the inverse modeling approach where our aim is it to find those genes that, by fixing them at a certain expression level, affect the model such that the generated artificial microarray dataset shows specifc traits. As described in section 2.2.4 we estimate the probability $P(C|E)$ of the incidence of a certain cancer-subtype C given some observation E, in this case, the expression-state of a certain gene $P(C|gene_i{=}state)$. A high probability predicts the fixed genes to be a potential cause for the subtype-specific expression-behavior of the queried genes that in turn can be the underlying reason for a specific cancerous phenotype.

For the comparision we used 7 reference-datasets where each of them arises from patients with a specific ALL-subtype. Figure 3 a) shows that the original microarray-dataset is clearly subdivided into 7 clusters of different sample-sizes. Each of this clusters represents the expression-pattern of 271 genes given a certain leukemia-subtype and was used to measure the impact of evidence on the appareance of these different ALL-subtypes. In a first step we fixed each gene at any of its expression-values using each of this conditions to generate a dataset of 300 samples. We then compared each of this data with the 7 reference-datasets as explained in section 2.2.4. In Figure 4, the probability of each subtype given one gene is overexpressed

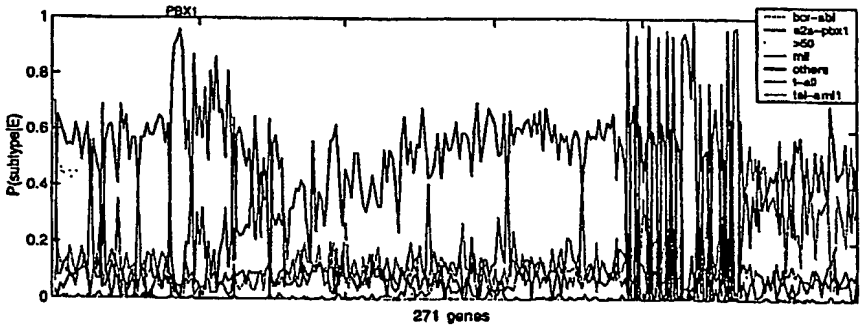


Figure 4: The probability $P(subtype|{+}1)$ for each subtype given gene i is overexpressed. For some genes a subtype-specific pattern appears with a probability near to 1.

is plotted over all 271 genes. Apparently, there exist a small number of genes, that evoke a certain ALL-subtype with a high probability, given they are highly active. To proof our results we will have a closer look at the molecular function of certain genes and their role in biological processes especially regarding pathogenesis.

Fig. 6

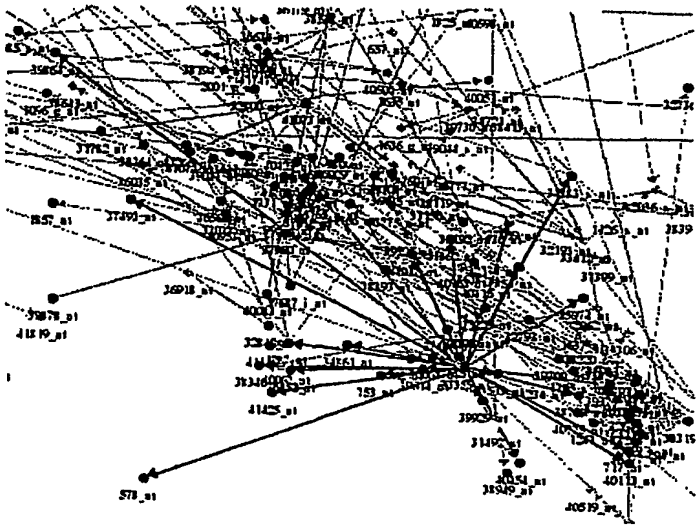


Figure 5: The dominant character of PBX1 confirms that it works as a transcriptional activator.

3.3 Biological insights

For this examination we will look into genes that have a high probability to cause a certain subtype as well as into significant structural pattern in the learned network, e.g. dominant genes and their surrounding. The learned Bayes-net results from a microarray dataset of different leukemia-subtypes and reflects transcriptional relationships among genes that occur in these malignant cancer-cells. Thus, genes that causes a certain subtype should either be potential oncogenes or regulated by such genes.

The first gene that we want to analyze more specifically is gene PBX1. When it is overexpressed our model generates with a probability of 0.96 a dataset that is characteristic for ALL B-lineage subtype E2A-PBX1 (cf. Figure 4). This suggests, that there exists a causal relationship between the overexpression of this gene and the incidence of ALL subtype E2A-PBX1. And in fact, PBX1 is known as an proto-oncogene causing the transformation of normal blood cells into malignant ALL cancer-cells. Due to the chromosomal translocation t(1:19), PBX1 fuses with gene E2A and converts to a potent oncogene causing leukemia subtype E2A-PBX1 []. Furthermore, since the graph structure of the model can be interpreted in a causal manner it gives information about the interaction between potential oncogenes and other ones which in turn can be interpreted as an oncogenic regulation. Looking

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☑ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☑ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☑ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☑ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.